Environmental Health Perspectives Supplements
101 (Suppl. 2): 145–152 (1993)

Design Options and Methodological Fallacies in the Studies of Reproductive Failures

by Jørn Olsen[1] and Torsten Skov[2]

Reproductive failures are at first sight well suited for epidemiologic research. The time of pregnancy is closely monitored, and failures such as spontaneous abortions and subfecundity are rather frequent. Although epidemiologists' interest in the field has been growing, there is still disappointingly little new information of relevance for prevention. A number of methodologic shortcomings may explain this. A large part of disease classification is not well suited for etiologic research, reduced fertility has diminished the populations at risk, close medical monitoring tends to mask causal links, and many scientific problems related to this area bring limitations to the research field. Still, much more could be learned from a systematic use of epidemiologic knowledge, existing registers, and the joint effort between different research groups.

Introduction

Reproductive failures in humans are common and are of many different types. Reproductive failures may be related to failure to conceive or failure to achieve a successful outcome of pregnancy. Reproductive failure may be prevented if the cause is known. The present state of our knowledge makes it possible to prevent only a small fraction of all reproductive failures; however, an increasing number of studies have supplied us with new information concerning environmental, occupational, and lifestyle factors that may adversely affect reproductive outcome. Reproductive failure would seem to be well suited to epidemiological research, and this subject has now attracted the interest of a number of epidemiologists with the prospect of an even greater expansion of our knowledge.

Epidemiological investigation of reproductive failures may be expected to be fruitful in the short term because pregnancy lasts for only a short time period and is covered by intensive medical surveillance. Numerous medical examinations are mandatory and these are performed with a high degree of standardization. The data are usually recorded and frequently stored in computer files. There is normally accurate registration of birth data so that the date of the pregnancy can be correlated with potential exposure to risk factors, and there is plenty of opportunity to standardize the collection of relevant information. Exposure data may be compiled during the course of the many health examinations which otherwise may be of dubious value.

Pregnancy is a major event in most people's lives, and the events that occurred during pregnancy are often accurately recalled even several years afterwards. In addition, reproductive failure is a frequent occurrence. Therefore, there is a potential for ideal research circumstances. However, closer scrutiny identifies a number of problems and shortcomings that may explain why new knowledge is only coming to light very slowly. A number of these problems are related to the specific types of outcomes studied, ranging from low fecundity to first- or even second-generation effects. Nevertheless, some problems are of a more general nature and are relevant to the study of several different types of outcome.

The population at risk for experiencing a reproductive failure is not large when the overall fertility or pregnancy rate of that population is generally low. A large proportion of a population may be exposed to chemicals, physical agents, or lifestyle circumstances that have an adverse effect on reproductive outcome, but this will be of little relevance if they have no desire for having a child or further children. Such effects are not detectable in studies of the general population unless exposure to noxious factors becomes manifest as biological changes such as semen quantity or quality, hormonal disturbances, or alterations in sexual function. A prerequisite for pregnancy usually involves unprotected intercourse at the very least, which may, or may not, be associated with a desire for having a child. Many social, cultural, occupational and other factors

[1] The Steno Institute of Public Health, Department of Social Medicine, University of Aarhus, Høegh Guldbergs Gade 8, 8000 Aarhus C, Denmark.

[2] Danish Cancer Registry, Institute of Cancer Epidemiology, Danish Cancer Society, Rosenvængets Hovedvej 35, Box 839, 2100 København Ø, Denmark.

This manuscript was presented at the Conference on the Impact of the Environment on Reproductive Health that was held 30 September–4 October 1991 in Copenhagen, Denmark.

affect the desire for having children or the lack of consistent use of birth control, and these may also be influenced by past reproductive history. All these factors may lead to self-selection for which it is difficult to adjust in the design of the study or in the analysis of the data.

Self-Selection

Bjerkedal and Erickson have clearly shown that birth outcome has consequences for subsequent fertility (*1*). If the first child dies, subsequent pregnancy occurs much more frequently than if the child survives, and even the sex of the first and second children may play a role in subsequent family planning. Couples with two children of different sex are less likely to have a further pregnancy than couples with two children of the same sex.

The likelihood of subsequent pregnancy is also affected by other outcomes of previous pregnancies. A miscarriage is likely to be followed by a further pregnancy if the child was wanted. However, a severe congenital malformation in a baby may have a profound influence on the desired family size(*2*). Social, economic, and personal factors also influence the desire to achieve pregnancy. Although external factors that influence fertility may be controlled in the design of studies or in their analysis, it is much more difficult, or even impossible, to control forces of self-selection if these are related both to the exposure to the agent under investigation and to the underlying risk of reproductive failure.

Studies of fetal loss usually adjust for pregnancy order or gravidity in the analysis, especially if the number of previous pregnancies differs among the groups to be compared. This procedure is probably rather straightforward if pregnancy order in itself carried a risk of reproductive failure: vital statistics indicate that such a risk exists. Most types of reproductive failure vary in frequency according to order of pregnancy, the graphic demonstration often being U-shaped, with the lowest risk being associated with the second pregnancy. However, surveys of women with a defined final number of pregnancies give a different picture. Backwards recording of such women indicate a reduction in the risk of reproductive failure in successive pregnancies, and the highest starting risk is among those women who have many pregnancies during their fertile period (*3–5*).

Assuming that a population under study is homogeneous in its desired family size, it can be concluded from the above data that fetal losses are compensated for by new pregnancies. Consequently, good reproducers retire from childbearing earlier than poor reproducers. If the groups to be compared desire to have the same size of family, then gravidity could be a proxy index of the women's inherent risk of fetal loss. Furthermore, pregnancy order would also seem to be a proxy index of some inherent risk of fetal loss that diminishes with each subsequent pregnancy. The association between pregnancy order and reproductive failure may, in fact, be a result of the self-selection process (*6*) because the outcome of pregnancy may be entirely dependent on the desired family size and therefore may have no bearing on biological influences. If specific cases are considered individually, a single case of fetal loss may be compensated for by a new pregnancy, whereas the birth of a handicapped child may delay future pregnancies. It is possible, therefore, that pregnancy order is only a proxy variable for a true potential confounding factor. If the groups to be compared desire the same family size, then pregnancy order may be a true measurement of an unknown confounding factor related to the risk of reproductive failure, although pregnancy order itself may only be a proxy variable. If that is not the case, then it is unlikely that the proxy variable would be able to adjust for the confounding factor. A better choice, if it were possible, would be to adjust for maternal age and the desire for a given family size. If appropriate groups are selected for comparison, a desire for a similar family size will, under the null hypothesis, produce unbiased results.

The only satisfactory solution to the statistical problem of lack of independence, and the scientific problem of the influence of previous reproductive history on the present desire for a new child, is to restrict the study to include only primigravidas. However, this could seriously reduce sample sizes, and in most studies only parity data are available. In practice the only option is to study primiparae, which does not solve all problems related to self-selection based on past reproductive experience.

Adjusting for Gravidity and Parity

At present, there is no fully satisfactory general rule that applies to the handling of past reproductive history in nonrandomized studies. However, simulation models have been devised in which the true risk of reproductive failure in relation to past pregnancies is expressed in terms of parity and gravidity. Such simulation models show that specific data, adjusted at a more crude level, will usually be quite close to unbiased values of measurements of effects if the groups studied are comparable in respect to desired family size (*7*). If exposure to certain factors has no effect on reproductive outcome, then it is likely that making comparisons of groups which are expected to have comparable desired family size will produce unbiased results, on the assumption that no other sources of bias are in operation.

The study period can be expanded to cover more than one pregnancy in given individuals, and this would appear to be an elegant way of allowing each woman to be her own control by comparing pregnancies which are exposed to various factors with previous or subsequent unexposed pregnancies. The rationale behind this approach presupposes that to some extent pregnancies occur independently of each other and have their own pattern of exposure to different agents. If parity itself is a risk factor, then the study base would have to be balanced by taking parity into consideration, which may be difficult or even impossible to achieve. The design of a study is also less attractive if social factors associated with exposures, or the exposure itself, have any influence on the planning of a pregnancy in the light of previous pregnancy history. Unfortunately, this will often be the case.

Time and Independence

Some studies may include data culled from multiple pregnancies in some individuals, and this has a bearing on another problem that is often exaggerated by many peer reviewers. If a given study includes some women who have more than one pregnancy under investigation, these pregnancies are not independent observations in every respect and perhaps should not be treated as independent observations in a given statistical evaluation. It is true that some causes of reproductive failures, such as genetic factors, chronic diseases, etc., are common background factors for all pregnancies in the course of a woman's reproductive life and these background factors have a bearing on self-selection due to previous pregnancy history. However, in most instances each new pregnancy can be treated as being a new experiment with its own potential hazards. The statistical error due to lack of independence between pregnancies will give rise to more narrow confidence limits than is strictly justified when all pregnancies are used in the study, but this error is small in most cases (*8*). However, habitual aborters may cause quite misleading results in small studies and there should always be a check to make sure that a given association is not due to a few outliers who have extremes of pregnancy outcomes.

Pregnancy requires the sexual relationship of two different individuals and this makes the timing of relevant exposure to various factors difficult to assess, because not only may the exposure of the female partner during pregnancy be of importance, but the exposure of the male and/or female partners before pregnancy may also be of relevance (*9*). Theoretically, any exposure of either parent prior to conception could have an effect, and this includes exposure during their fetal life or even in previous generations. In most cases, exposures of most relevance probably involve the male partner during spermatogenesis and the female partner around the time of conception and during pregnancy. Epidemiology usually involves studies of individuals, but the epidemiology of reproductive failures involves the study not only of individuals but also of couples, and in most cases the additional investigation of one or more fetus(es). A further complication when designing a study is to take into consideration a society which has a possible rapid turnover of partner relationships.

Specificity and Classification

Lack of specificity of effect is sometimes to be expected. A low intake of alcohol may have no effect on brain development, or it may have a minor effect. A high exposure to alcohol may cause severe mental retardation, birth defects, spontaneous abortion, or infertility. This lack of specificity, therefore, makes case–control studies less attractive because the case–control approach may only be able to detect exposures of a certain timing or intensity. This can only be circumvented by applying a series of case–control studies with different case groups according to different reproductive failures.

Some of the more frequent measurements of outcome are very crude, such as the incidence of spontaneous abortion which is the most popular reproductive failure to be studied. This is a measurement of general mortality, but in most cases exposure to adverse factors does not cause such an extreme outcome unless the severity of the exposure is sufficient to result in cause-specific mortality. Therefore, spontaneous abortion as a measurement of outcome is not as attractive as it would first appear (*10*). It should be borne in mind that the amount of information that can be generated by a study does not depend on the frequency of exposure in a population and the specific outcomes in that population, but depends on only the frequency of outcomes related to the specific exposure. If it is not possible to separate relevant outcomes from irrelevant outcomes, the results will be serious misclassifications which will bias the risk estimate toward null values.

Exposure to vinyl chloride has very little impact on general mortality, but is strongly associated with a particular type of liver cancer mortality (*11*). Likewise, exposure to many different agents may be strongly related to specific types of abortion, but these would not be detected in studies of abortion in general. The same is true for studies on infertility or fecundity, fecundity being the probability of conception within a given menstrual cycle when there is normal sexual cohabitation without the use of contraceptives. Subclassification of subfecundity or infertility will most likely increase the amount of statistical and scientific information which can be gained from a study, but unfortunately such data are not normally readily available because only a small fraction of subfertile/infertile couples seek medical help (*12*). The different types of abortion may be subclassified according to clinical and genetic characteristics, but this is a difficult and expensive exercise.

Prevalence or Incidence of Reproductive Failures

The preferred methods of inferring causal relationships are usually to assess incidence rates or to estimate relative risks, but these are rarely used for the study of reproductive failures. It is usually not possible to follow the fate of embryos from their conception and to record relevant outcomes as a function of observation time or number of embryos/fetuses at risk. In most cases the only embryos to be studied are those which survive until pregnancy is clinically recognized, or until birth. Therefore, in a study of birth defects, only prevalence data can be recorded because only the prevalence of malformations which survive until birth can be detected. The prevalence rate is not a measure of the total number of cases, but is only a measurement of a proportion of cases: it is the proportion of a population that has the disease at a specific point in time, and in the example of birth defects, the specified point in time is the moment of birth.

Exposure to a substance may be instrumental in causing a malformation, or it may have the effect of preventing the spontaneous abortion of a malformed fetus: in either case, such a substance will increase the prevalence of malformations at birth. Exposure to a substance which induces the mortality of a malformed fetus can cause a reduced prevalence of malformations at birth, even if that substance has a causal effect in increasing the true incidence of that malformation. This elementary fact has been known for years (*13*), and it is therefore difficult to understand why the term "incidence of malformation at birth" still appears in the toxicological literature. Studying events at birth is to study perhaps a small, and usually biased, sample of potential reproductive failures, and although the adverse effects may have occurred months previously, only those surviving till birth are examined.

The intensive screening activities that take place during pregnancy by the health care system are related to the problem of conflicting incidence and prevalence rates. Studies of birth defects refer to reproductive failures which are not removed naturally during pregnancy prior to birth and are not prevented or terminated by health care intervention. Of particular interest is the screening for congenital malformations, the interpretation of which requires very reliable data on whether the malformations are detected at birth or in a prenatal screening program. Prenatal screening could heavily bias the results toward a higher recorded frequency and groups to be compared with each other must be stratified according to screening activities.

There are, of course, other methodological problems in designing epidemiological surveys, particularly in the investigation of reproductive failures, but most of these are related to the study of specific types of exposure or specific outcomes. Some of these will be mentioned in the following section, which is devoted to frequent and often studied reproductive failures.

Infertility/Subfecundity

Infertility is one of the more frequent reproductive failures. Many couples have greater than expected difficulty in conceiving or having a child. About 15% of couples fulfill the criteria for infertility, which is defined as no pregnancy after at least 12 months of trying to conceive (*14*). Fecundity, the probability of conceiving within a given menstrual cycle, is probably 0.30 on average in Denmark, but variations are large between couples and perhaps also within couples over time.

In principle, fecundity is rather easy to study: couples merely have to start counting cycles or the time to pregnancy after they stop using contraceptives. The number of cycles or months until pregnancy directly relates to fecundity, and such a measurement can be used to compare groups which are characterized by different types of exposure under investigation. Couples who have a prolonged time to pregnancy are likely to have low fecundity, but this is not always true because conception is a random phenomenon with a specific chance of success. Therefore, some couples may have to wait for a long time purely because of bad luck, whereas some subfecund couples may have the good luck of conceiving within a short period of time.

The probability of no pregnancy within six cycles in spite of a normal fecundity of 0.30 is $(1-0.3)^6$, or 0.12, and within 12 cycles is $(1-0.3)^{12}$, or 0.01. Because the latent fecundity is only measurable by means of a waiting time, some misclassification of subfecundity at the individual level is unavoidable, regardless of the means of categorization. If subfecundity is defined as at least six months of trying, 12% of couples with normal fecundity will be classified as having some degree of subfecundity. On the other hand, after 12 months, some subfecund couples will be classified as fecund; for example, a couple with a fecundity of, say, 0.10 have a probability of conception within 12 months of $1-(1-0.10)^{12}$, which is 0.72. This unavoidable misclassification has clinical relevance, but is of less concern in population studies. The proportion of fecund couples in a population is measurable by recording total numbers in that population; however, the effects of exposure to different factors will be biased toward their null values. From a statistical point of view, it is better to compare the distribution of waiting time to pregnancy or number of cycles to pregnancy (*15*). However, experience shows that reliable data are difficult to obtain from short periods of waiting time.

The main problem about using measurements of waiting time is that pregnancies have to be planned. Most women can rather accurately recall the waiting time, even for pregnancies which occurred years previously, if the pregnancy was planned and if reliable methods of birth control were used until then (unpublished results from piloting in the European Studies on Infertility). Problems arise if pregnancies occur in spite of the use of contraceptives, especially after irregular use of contraceptive methods. Women who become pregnant in spite of regular use of contraceptives are probably fecund and may be classified as such, but it is more difficult to classify couples who achieve pregnancy while using contraceptives irregularly. Some couples from their own experience estimate that they have low fecundity and therefore use contraceptives irregularly or not at all. Procedures such as surgical sterilization remove couples from the population at risk and must not be counted. Studies have shown that the proportion of couples who have had normal sexual relationships without the use of contraception for at least 1 year is more than twice as high as the proportion of couples who have tried to become pregnant for at least 1 year (*16*).

In countries where the use of contraception is generally irregular, or unsafe methods are used, many people probably have a good idea of their own fecundity based on their own sexual experience: this causes major problems when making international comparisons of infertility. Calculating infertility rate based on the lack of success of achieving a desired pregnancy for at least 12 months is quite different from making calculations based on at least 12 months of unprotected intercourse. This makes it necessary to decide whether or not the study base should be restricted to couples who are trying to achieve pregnancy. If so, pregnancies have to be planned, and fortunately the majority of pregnancies are now planned in some countries. Valid comparisons may still be possible in many cases if careful adjustments for contraceptive habits are used in the analyses.

The most straightforward design of a fecundity study is a follow-up study. However, if only contemporary data are recorded, the source population usually has to be large. Alternatively, a full pregnancy history may be collected and a validated questionnaire has been developed for such purposes in Europe (European Studies of Infertility).

A cross-sectional recording of subfecundity will, of course, be heavily biased toward the longest waiting times, which occur in sterile couples. This problem must be taken into consideration if the level of exposure under study changes with time, to counteract bias in the data analysis.

The recording of fecundity is a crude measure of reproductive outcome because of the many different potential causes. It will detect not only failure of conception but also early spontaneous abortion, which is not recognized clinically. Some early abortions can be differentiated from lack of conception by close surveillance of pregnancy hormones, but this is an expensive option. Other possibilities for detecting narrower subsets of subfecundity include semen analysis and other standard investigations of infertility. The narrower subsets will have fewer potential causative factors, and this will increase the possibility of detecting quantitative relationships. However, this is also expensive and difficult to implement and should perhaps not be used as the first choice in larger studies.

Exposure to most reproductive toxins, including dibromochloropropane (DBCP), causes subfecundity and only results in sterility after very heavy exposure. For substances that can cause either sterility or reduced fecundity, depending on level of exposure, an unusual case–control option is available (*17*). Individuals with a long waiting time to pregnancy can be easily identified in countries with centralized facilities for surveillance of pregnancies or deliveries by means of a short questionnaire or interview; a standard questionnaire produced by the European Studies of Infertility is available. Information from the questionnaire could be used to compile a study base restricted to couples who used reliable methods of birth control, and cases identified as being subfecund could be compared with a sample from the study base. By applying restrictive case–control sampling it may be possible to subclassify subfecundity into various clinical subsets by offerring a medical examination to both of the prospective parents. If common lifestyle factors such as smoking, coffee drinking, etc., are studied, it is advisable to restrict the study base further to primiparas only.

The average waiting time to pregnancy will be prolonged if infertile couples are included in the group, therefore, infertility should be an exclusion factor for a study in order to avoid selection bias. However, if cases of subfecundity are identified from a defined population which includes all pregnant women who have reached a given gestational age, selection bias is avoided because infertility does not need to be considered as a selection problem because infertile couples will automatically be excluded from the test and control groups. Such a study would, of course, not be able to identify the effect of exposure to a substance if the exposure level under consideration could cause sterility without having any effect on fecundity amongst those couples who remained fertile. This scenario, however, is rare.

Case-sampling taken from patients seeking help in sterility clinics carries a high risk of selection bias. It is known that 30–50% of infertile couples do not seek help (*12*), although this proportion may change depending on the development of better clinical facilities. Housing conditions, other social factors, and probably also school education may influence whether or not individuals seek advice (*12*). Occupational exposure to a specific substance may also affect the decision to seek medical help if there is a suspicion that that substance is a reproductive hazard. One possibility for avoiding such selection problems would be to select controls from within the same group of patients attending the clinic for sterility problems. The effect on male fertility can be examined by identifying exposed males attending the clinic who have objective signs of reduced fertility and comparing these with reference males who were similarly exposed and being investigated for infertility, but had no male medical problems diagnosed. However, attempts to verify findings from such studies using other studies with different designs have to some extent been disappointing (*18*), except for the association between exposure to welding and infertility (*19,20*). There is still limited experience in this field and more studies are to be welcomed.

The basic assumption behind designing a case–control study using patients from within the health care system is that the same forces of selection exist for all types of infertility, but this need not be the case. Males exposed to DBCP are probably more likely to be overrepresented in a group of patients seeking treatment for infertility than would be expected by their frequency among all males with a low sperm count. When studying substances which are suspected by those exposed, or by their doctors, to cause the specific effect under investigation (i.e., infertility), usual case–control techniques become problematic.

Standardized Fertility Ratios

In countries where registers are kept, all births are documented with computerized links to their biological mothers and to those who are classified as their fathers. Demographers use these data to estimate fertility rates, and measurements of fertility have also been used in epidemiology. The main problem is, of course, that reproduction does not primarily depend on biological performances, but rather on a number of social and cultural factors. Differences in fertility rates between groups need not necessarily reflect biological fecundity, although it may do so to some extent. In the early days of the DBCP investigations, a design for fertility studies was proposed which was claimed would solve some of the problems associated with using fertility as a measure of outcome (*21*). It was proposed to measure fertility as a standardized fertility ratio, which is the observed fertility in a given period of time divided by the expected fertility based on the actual fertility of couples of the same age from the same region. Interesting results may emerge from computing a standardized fertility ratio for a given cohort before, during, and after exposure to a substance under investigation. This method has been shown to detect the DBCP effect, but almost any method would have been capable of demonstrating the effect of this substance. The method has also been applied to a cohort of welders (*22*) and has given results comparable to other methods. It is a method which is likely to be useful as a screening tool, making use of record linkage applied to existing computerized data. More in-depth studies using case–control techniques could be performed if preliminary studies demonstrate a low standardized fertility ratio.

Spontaneous Abortions

Many pregnancies, perhaps more than 30%, end in a spontaneous abortion (*23*), but many of these abortions are not detected unless pregnancy is diagnosed using close hormonal surveillance. Among routinely recognized pregnancies, 8–20% end in abortion. Many studies have been devoted to the investigation of this outcome, but there is still disappointingly little known about their preventable etiology. Perhaps further scientific progress must await subclassification of abortions according to their clinical and genetic characteristics, but this is difficult and expensive to do.

In many countries, most recognized abortions are diagnosed and treated in the health care system, and in some countries the events are recorded on computerized files. In a number of studies, such outcome registers have been linked with union files to produce job-specific rates or ratios of spontaneous abortion. Past experience has shown, however, that more detailed data are needed to obtain relevant scientific information that could be useful for developing prevention strategies.

In the Nordic countries, almost all existing birth and abortion registers are used in epidemiological studies to establish a proper study base that includes a reasonable proportion of those exposed. An abortion register and birth register linked with, say, a register of members of a given union may be used to perform a case–control study within a cohort. Such a cohort would consist of embryos and fetuses that survived long enough *in utero* to produce a clinically detected pregnancy, and a register of abortions should include induced abortions. In relation to a cohort study, induced abortions are censored observations and could be excluded from the study if the reason for the termination of pregnancy is unrelated to the exposure under investigation. Life table analyses usually cannot be applied to spontaneous abortions because of differential left censuring, and this has a major influence on the validity of using artificial measurements on the rate of spontaneous abortions, especially the ratio of spontaneous abortions to births (*24*). Therefore, it is important to make sure that the frequency of induced abortions is similar in the groups to be compared or to select comparison groups that are likely to have the same proportion of induced abortions.

Simulation models show that the proportion of induced abortions introduces more bias than the timing of the induction (*24*). In most cases, a direct relationship between exposure and the frequency of induced abortions is unlikely, but an indirect association may be caused by social factors associated with the exposure. The frequency of induced abortions varies between different social groups according to urbanization, school education, etc. If the design of the survey controls for such factors, the timing and frequency of induced abortions are likely to be similar in the groups to be compared.

Detailed recording of exposure in relation to the gestation time of pregnancy is usually necessary, and exposure before pregnancy or in early pregnancy should be documented. In a case–control study, the controls should be pregnant women sampled from the same population at risk and matched for gestational age so that the timing of the exposure is likely to be the same in both groups. Unfortunately, this is rarely the case. More often, the affected individuals and the controls are selected after the events have taken place. The true risk of abortion, or abortion rate, cannot be estimated based only on routinely recorded abortions, and in such circumstances a ratio of abortions to births may have to be used as a measurement of the association between the exposure and subsequent abortion. The control women who progress to give birth to a child should match the study group in the time window during which the exposure occurred. If an exposure varies during the course of the study period, as for example occurs with the use of video display terminals (VDTs), the calendar time for the two groups should also match. If previous pregnancies in addition to the present pregnancy are used in the study, the problem of the exposure time is very pertinent, and it is important that the individuals can remember details of the time of exposure in relation to the gestational age. The timing between the present pregnancy and a previous abortion is likely to be shorter than the spacing between the present pregnancy and a previous birth.

The choice between a follow-up study or a case–control study mainly depends on the opportunities which are available for retrospective recording of exposures and on the problems related to recall bias. If details of a previous abortion are recorded in an interview that relies on recall of events that occurred some time in the past, it has been discovered that the accuracy of the recall may be poor after a period of 4 years have elapsed (*25*). It is often difficult to differentiate between a late onset of menstrual bleeding and an early abortion, and this uncertainty has the potential of introducing detection bias in a follow-up study.

Birth Weight/Fetal Growth

Much more is known about the determinants of fetal growth than any other measurements of pregnancy outcome, therefore, epidemiological studies on birth weight are numerous (*26*). Most of such studies have been of the follow-up type, using birth weight as the measure of outcome, but case–control studies have also been published using low birth weight babies as cases, the criterion for selection being less than 2500 g at birth. Birth weight is usually recorded according to rather standardized procedures of which the timing of the weighing of the baby after birth is of paramount importance.

The main problem concerning studies of birth weight is the fact that birth weight relates not only to fetal growth but also to gestational age. A low birth weight may be due to reduced fetal growth, low but normal fetal growth as an effect of genetic inheritance or preterm delivery. Retarded fetal growth may be differentiated from a genetically determined small baby by adjusting for parental height and birth weight of previous pregnancies and adjustment for gestational age can help to counteract for preterm delivery.

Many publications have tackled the problem of controlling for gestational age. A restriction may be applied to the study base so that only full-term deliveries are included, or the series may be stratified by gestational age in the analysis. Adjustment for gestational age has quite often been achieved by including gestation in the statistical model.

An alternative approach has been suggested whereby controlling for gestational age is substituted by the use of a statistical model. It is well known that birth weight varies with gestational age, and heteroscedasticy exists between gestational age and birth weight (*27*). To overcome such statistical problems, the use of a birth weight ratio measure has been suggested (*27*), The observed birth weight is divided by the expected birth weight according to the given gestational age. A population register of birth weight at different gestational ages is used to estimate the expected values, and the ratio is used in the model to replace birth weight as the measurement of outcome. The main drawback to this approach is related to the problem of the reliability of assessing gestational age. Furthermore, this model uses an external reference group that is taken into the statistical analysis, but details of exposures in this reference group are unknown and uncontrolled.

Another approach has been suggested which takes into consideration the principles of a confounder score (*28*). An estimate of expected birth weight is made on the basis of data derived from the nonexposed individuals in the study with adjustments being made for all confounding factors. In doing this, it may be necessary to exclude births with a short gestational age because numbers may be too few to make reliable estimates (*29*). A simple outcome measurement is then obtained, some statistical problems are solved, and adjustments are made for confounders in the normal way.

In some circumstances, it may be wise to avoid any combined measurement based on gestational age. Often the data on birth weight are much better than the data on gestational age, and a composite measurement may turn good data into bad data. It may be better to study only birth weight, but in such circumstances this restriction must be taken into consideration when analyzing the data and discussing the results because of the possibility that low birth weight may be a direct result of preterm delivery.

Many of the screening procedures that take place during pregnancy are for the identification of fetal growth retardation, and elective Cesarean section may mask severe growth defects. Multiple births are also usually associated with fetal growth alterations, and it is normally only singleton births that are used for studying the determinants of birth weight.

There is usually a normal distribution of birth weight with a rather symmetric pattern, and if exposure to a certain factor has a tendency to alter the median or mode of the distribution without modifying its shape, then the birth weight should be used as the measurement of choice for detecting the effect. However, should exposure to a given substance modify the shape of the distribution rather than its mean, then a different end point should be measured. A time-honored definition of low birth weight is 2500 g or less, but even at term this cut-off value would include some newborns who had a birth weight appropriate for their genetic constitution. The proportion of newborn babies who are of low birth weight by definition but are genetically small depends on the population under study. Other definitions of normal size depend on deviations from the normal distribution of birth weights according to gestational age. Two problems are related to this, one being the origin of the data used for calculating the normal distribution and the other being the measurement of deviations from the normal distribution using standard deviations or percentages.

In comparing birth weight distributions, Wilcox and Russell have advocated the comparison of the shapes of the distribution rather than their central tendencies (*30*). They observed that indexes such as perinatal mortality were more closely associated with the deviation from the normal distribution than the mean birth weight. Furthermore, they have developed computer software that fits the best possible Gaussian distribution to the data available and calculates the proportion of cases falling outside the left-hand margin of this distribution curve. According to Wilcox and Russell, this residual fraction is more predictive of reproductive failure than the mean birth weight, but this need not be true for all populations.

Another controversy related to similar problems has been the comparison of specific reproductive failures, stratified by birth weight. Even when the problems of adjusting birth weight for the variables that may have an influence and that may also be intermediates in the causal links under study are taken into consideration, birth weight itself is still only a proxy variable for the potential confounding factors that would be desirable to control. The correlation between the proxy variable and the true confounding factors may vary from population to population, depending on all the causal factors affecting fetal growth, on the reasons for preterm and post-term deliveries, and on the genetic coding for growth.

Congenital Malformations

Congenital malformations have been the subject of several studies, especially since the thalidomide tragedy. Some malformations are serious and easy to diagnose, and although most of the specific malformations are rare, they may be obvious candidates to study in comparison to controls. Many case registers have been set up for this purpose, but only a few scientific reports have been published, and perhaps there are too few cases to justify the costs. With surveillance now in operation, another potential thalidomidelike disaster should not be missed. However, other teratogens may be overlooked because of lack of statistical power in most surveillance systems.

Limited resources for research are not the only reason for the possibility of overlooking potential teratogens. There are also a number of scientific limitations. The first limitation relates to the measurement of prevalence rather than incidence. Only malformations in fetuses that survive until birth are normally eligible for study, and many malformations are part of syndromes that are not compatible with fetal life. The second limitation is related to diagnostic problems. Registers often have to be used, and the quality of the register depends heavily on the clinical setting. The prevalence rate of malformations could vary from 0.01 to 0.08 depending on the degree of specialization of the physician responsible for the clinical case identification. The third limitation relates to a disease classification that is based on certain principles but not on principles associated with causal research.

There is still much more research that could be done on congenital malformations. Many of the large registers such as EUROCAT have not been used to their full potential. There are still unused opportunities in international collaboration. Epidemiologic studies could make use of measurements of biological exposure, and toxicologic studies should make more use of epidemiologic techniques.

Other Reproductive Failures

There are a number of other measurements of reproductive outcome that could be used such as gestational age, the sex ratio of abortuses/children, twinning rates, Apgar scores, early deaths, etc. In studies on the use of alcohol and tobacco during pregnancy, long-term effects have been investigated, and of major interest are the effects not only on childhood cancers, but also on cancers developing later in life such as testicular cancers and on allergies and mental development. The opportunities for such long-term studies have not been developed as much as they should.

Studies on twins have shown one of the best-documented changes in reproductive outcomes. Twinning rates in most countries, including Denmark, dropped by more than 20% over 5 decades, mainly due to a decline in dizygotic twinning (*31*). There are still no good explanations available to account for this secular trend, and the increasing use of hormonal treatment for subfecundity has reduced the options for further research into this issue.

REFERENCES

1. Bjerkedal, T., and Erickson, J. D. Association of birth outcome with subsequent fertility. Am. J. Obstet. Gynecol. 149: 399–404 (1983).
2. Record, R. G., and Armstrong, E. The influence of the birth of a malformed child on the mother's further reproduction. Br. J. Prev. Soc. Med. 29: 267–273 (1975).
3. Bakketeig, L. S., and Hoffmann, H. J. Perinatal mortality by birth order within cohorts based on sibship size. Br. Med. J. 2: 693–696 (1979).
4. Roman, E., Doyle, P., Beral, V., Alberman, E., and Pharoah, P. Fetal loss, gravidity, and pregnancy order. Early Hum. Dev. 2: 131–138 (1978).
5. Billewicz, W. Z. Some implications of self-selection for pregnancy. Br. J. Prev. Soc. Med. 27: 49–52 (1973).
6. Mantel, N. Perinatal mortality by birth order. Br. Med. J. 2: 1147 (1979).
7. Olsen, J., and Heidam, L. Z. Analysis of pathological outcome of pregnancy. Scand. J. Soc. Med. 11: 3–6 (1983).
8. Butler, W. J., and Kalasinski, L. A. Statistical analysis of epidemiologic data of pregnancy outcomes. Environ. Health Perspect. 79: 223–227 (1989).
9. Selevan, S. G. Design of pregnancy outcome studies of industrial exposures. In: Occupational Hazards and Reproduction (K. Hemminki, M. Sorsa, and H. Vainio, Eds.), Hemisphere Publishing Corporation, Washington, DC, 1985, pp. 219–229.
10. Olsen, J. Methodological problems in the studies of reproductive failures. Scand. J. Soc. Med. 16: 217–221 (1988).
11. Spirtas, R., Beebe, G., Baxter, P., Dacey, E., Faber, M., Falk, H., van Kaick, G., and Stafford, J. Angiosarcoma as a model for comparative carcinogenesis. Lancet ii: 456 (1983).
12. Rachootin, P., and Olsen, J. Social selection in seeking care for reduced fecundity among women in Denmark. J. Epidemiol. Commun. Health 35: 262–264 (1981).
13. MacMahon, B., Pugh, T. F., and Ipsen, J. Epidemiologic methods. Little, Brown and Company, Boston, 1960.
14. Olsen, J., and Rachootin, P. Prevalence and socioeconomic correlates of subfecundity and spontaneous abortion in Denmark. Int. J. Epidemiol. 11: 245–249 (1982).
15. Boldsen, J. L., and Schaumburg, I. Time to pregnancy—a model and its application. J. Biosoc. Sci. 22: 255–262 (1990).
16. Marchbansk, P. A., Peterson, H. B., Rubin, G. L., Wingo, P. A., and The Cancer and Steroid Hormone Study Group. Research on infertility: definition makes a difference. Am. J. Epidemiol. 130: 259–267 (1989).
17. Rachootin, P., and Olsen, J. The risk of infertility and delayed conception associated with exposures in the Danish workplace. J. Occup. Med. 25: 394–402 (1983).
18. Schaumburg, I., and Olsen, J. Time to pregnancy among Danish pharmacy assistants. Scand. J. Work Environ. Health 15: 222–226 (1989).
19. Bonde, J.P. Semen quality and sex hormones among stainless steel and mild steel welders: a crossectional study. Br. J. Ind. Med. 47: 508–514 (1990).
20. Bonde, J. P. Semen quality among welders at follow-up after three weeks of non-exposure. Br. J. Ind. Med. 47: 515–518 (1990).
21. Starr, T. B., and Levine, R. J. Assessing effects of occupational exposure on fertility with indirect standardization. Am. J. Epidemiol. 118: 897–904 (1983).
22. Bonde, J. P., Hansen, K. S., and Levines, R. J. Fertility among Danish male welders. Scand. J. Environ. Health. 16: 315–322 (1990).
23. Wilcox, A. J., Weinberg, C. R., O'Connor, J. F., Baird, D. D., Schlatterer, J. P., Canfield, R. E., Armstrong, E. G., and Nisula, B. C. Incidence of early loss of pregnancy. N. Engl. J. Med. 319: 189–194 (1988).
24. Olsen, J. Calculating risk ratios for spontaneous abortions: the problem of induced abortions. Int. J. Epidemiol. 13: 347–349 (1984).
25. Heidam, L. Z., and Olsen, J. Self-reported data on spontaneous abortions compared with data obtained by computer linkage with the Hospital Registry. Scand. J. Soc. Med. 13: 159–163 (1985).
26. Kline, J., Stein, Z., and Susser, M. Conception to Birth. Epidemiology of Prenatal Development. Oxford University Press, Oxford, 1989.
27. Bland, J. M., Peacock, J. L., Anderson, H. R., Brooke, O. G., and de Curtis, M. The adjustment of birth weight for very early gestational ages: two related problems in statistical analysis. Appl. Statist. 39: 229–239 (1990).
28. Miettinen, O. S. Stratification by a multivariate confounder score. Am. J. Epidemiol. 104: 609–620 (1976).
29. Olsen, J., and Olsen, S. F. A suggestion for improving intelligibility in multivariate confounder adjustment using birth weight as an example. A 'confounder score' approach in analyzing continuous data. Scand. J. Soc. Med. 19: 235–241 (1991).
30. Wilcox, A. J., and Russell, I. T. Birthweight and perinatal mortality: III. Towards a new method of analysis. Int. J. Epidemiol. 15: 188–196 (1986).
31. Rachootin, P., and Olsen, J. Secular changes in the twinning rate in Denmark 1931 to 1977. Scand. J. Soc. Med. 8: 89–94 (1980).